Bayesian classification of vegetation types with Gaussian mixture density fitting to indicator values.

Witte, J.P.M.; Wojcik, R.B.; Torfs, P.J.J.F.; de Haan, M.W.H.; Hennekens, S.

published in
Journal of Vegetation Science
2007

DOI (link to publisher)
10.1111/j.1654-1103.2007.tb02574.x

document version
Publisher's PDF, also known as Version of record

Journal of Vegetation Science 18: 605-612, 2007

Bayesian classification of vegetation types with Gaussian mixture density fitting to indicator values

Witte, Jan-Philip M.[1,2*]**; Wójcik, Rafał B.**[3]**; Torfs, Paul J.J.F.**[4]**; de Haan, Martin W.H.**[1] **& Hennekens, Stephan**[5]

[1]*Kiwa Water Research, P.O. Box 1072, 3430 BB Nieuwegein, The Netherlands;*
[2]*Vrije Universiteit, Institute of Ecological Science, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands;*
[3]*School of Engineering and Applied Science, Princeton University, C326 Engineering Quadrangle, Princeton, NJ 08544, USA;*
[4]*Wageningen University, Nieuwe Kanaal 11, 6709 PA Wageningen, The Netherlands;*
[5]*Alterra, P.O. Box 47, 6709 AA Wageningen, The Netherlands;*
[*]*Corresponding author; Fax +31 306061165; E-mail flip.witte@kiwa.nl*

Abstract

Question: Is it possible to mathematically classify relevés into vegetation types on the basis of their average indicator values, including the uncertainty of the classification?

Location: The Netherlands.

Method: A large relevé database was used to develop a method for predicting vegetation types based on indicator values. First, each relevé was classified into a phytosociological association on the basis of its species composition. Additionally, mean indicator values for moisture, nutrients and acidity were computed for each relevé. Thus, the position of each classified relevé was obtained in a three-dimensional space of indicator values. Fitting the data to so called Gaussian Mixture Models yielded densities of associations as a function of indicator values. Finally, these density functions were used to predict the Bayesian occurrence probabilities of associations for known indicator values. Validation of predictions was performed by using a randomly chosen half of the database for the calibration of densities and the other half for the validation of predicted associations.

Results and Conclusions: With indicator values, most relevés were classified correctly into vegetation types at the association level. This was shown using confusion matrices that relate (1) the number of relevés classified into associations based on species composition to (2) those based on indicator values. Misclassified relevés belonged to ecologically similar associations. The method seems very suitable for predictive vegetation models.

Keywords: Environmental Impact Assessment; Modelling; Phytosociology; Predictive vegetation model.

Abbreviations: *iv* = mean indicator value; pdf = probability density function; P = occurrence probability; C_k = vegetation type; GMM = Gaussian Mixture Model.

Introduction

In pursuit of general applicable relationships between vegetation and environment, ecologists often use plant traits or indicator values, in stead of individual plant species or vegetation types, as vegetation response variables (Diekmann 2003; McGill et al. 2006). Indicator values are often calibrated against physical or chemical field measurements of e.g. groundwater level, soil pH and soil nutrients (e.g. Diekmann 2003; Dzwonko 2001; Runhaar et al. 1997; Schaffers & Sýkora 2000). Such empirical relationships can be used to monitor environmental conditions from the vegetation or, *vice versa*, to predict vegetation from environmental conditions. Recently, Schmidtlein (2005) succeeded to map indicator values from airborne hyperspectral imagery.

In this paper we introduce a method to classify vegetation with the aid of indicator values or plant traits. This method enables us to transform maps of indicator values, obtained from predictive models or remote sensing images, into maps of vegetation types. Not only does our method classify vegetation types, it also yields the occurrence probabilities of vegetation types as a function of indicator values. It has already been applied in one model that can predict vegetation effects of water management, atmospheric deposition and vegetation management (Witte et al. 2004, 2006).

Vegetation types are usually delimited on the basis of the species composition of relevés and similar descriptions. In Europe it is common practice to classify the vegetation according to the syntaxonomical system of Braun-Blanquet. We will demonstrate that our method is capable of accurately distinguishing vegetation types at the association level. Moreover, we will show that it provides insight into the quality of both the indictor values and the vegetation classification system, as well as into the ecological requirements of vegetation types. Finally, we will argue why the method is very suitable to be used in predictive vegetation models.

Material and Methods

General approach

In the following four steps, occurrence probability functions of vegetation types were derived from a large database of vegetation relevés (Fig. 1):

1. Arithmetical mean indicator values (*iv*) were computed for each relevé. Following the findings of Käfer & Witte (2004), species abundance values were ignored.
2. Each relevé in the database was assigned to a vegetation type C_k on the basis of its species composition.
3. Thus, a secondary database was constructed, with C_k and *iv* values for each relevé, which is represented as a point in a *D*-dimensional space of *D* *iv*-axes. Relevés of the same vegetation type form a cluster in this space. Those clusters were described individually by probability density functions of indicator values. These functions may be interpreted as the 'ecological' envelopes of the vegetation types.
4. Finally, the calibrated density functions of all the vegetation types considered were used to predict the Bayesian occurrence probabilities of vegetation types as a function of indicator values: $P(C_k)$ = f(*iv*'s).

The occurrence probabilities thus obtained were used to classify relevés into vegetation types.

Bayesian classification and Gaussian mixture density models

Mathematically, the problem of classification of vegetation types based on indicator values can be solved in the classical Bayesian framework (Webb 2002). For *c* different vegetation types $C_1,\ldots,C_c$ and for a continuous vector $\underline{\mathbf{iv}}=[iv_1,\ldots iv_D]$ of *D* indicator values one can write Bayes' theorem in the form:

$$P\left(C_k\left|\underline{\mathbf{iv}}\right.\right)=\frac{p\left(\underline{\mathbf{iv}}\left|C_k\right.\right)P\left(C_k\right)}{p\left(\underline{\mathbf{iv}}\right)} \qquad (1)$$

Where $P(C_k|\underline{\mathbf{iv}})$ is the posterior occurrence probability

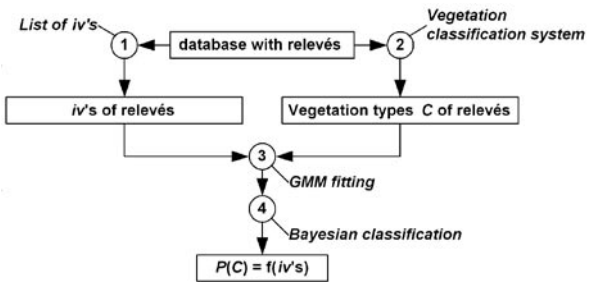

Fig. 1. Outline of our method to derive occurrence probabilities *P* of vegetation types *C* from indicator values *iv*. See section 'Material and Methods' for further explanation.

(the occurrence probability of vegetation type C_k once $\underline{\mathbf{iv}}$ is known), $p(\underline{\mathbf{iv}}|\ C_k)$ is referred to as the likelihood probability density function (pdf) for C_k and $P(C_k)$ is the prior probability of C_k. The unconditional density $p(\underline{\mathbf{iv}})$ is given by:

$$p\left(\underline{\mathbf{iv}}\right)=\sum_{k=1}^{c}p\left(\underline{\mathbf{iv}}\left|C_k\right.\right)P\left(C_k\right) \qquad (2)$$

which ensures that the sum of predicted occurrence probabilities:

$$\sum_{k=1}^{c}P\left(C_k\left|\underline{\mathbf{iv}}\right.\right)=1.$$

The prior probability $P(C_k)$ is simply taken as

$$N_k\Big/\sum_{k=1,c}N_k$$

where N_k is the number of relevés of vegetation type C_k.

To implement Eqs. 1 and 2, we need pdfs $p(\underline{\mathbf{iv}}|\ C_k)$ for all vegetation types C_k. One might model pdfs by parametric forms as e.g. Gaussian, lognormal or Gamma. In our method, however, we describe pdfs by nonparametric Gaussian Mixture Models (GMMs) which are defined as a linear combination of Gaussian densities (Fig. 2), called components:

$$p\left(\underline{\mathbf{iv}}\left|C_k\right.\right)=\sum_{j=1}^{J_c}w_j g_{\underline{\Theta}_j}\left(\underline{\mathbf{iv}}\left|C_k\right.\right) \qquad (3)$$

where *Jc* is the number of components and $g_{\underline{\Theta}_j}(\cdot)$ stands for the j^{th} Gaussian component defined by the set of parameters (means and covariances) $\underline{\Theta}_j$. Here w_j's are the component's weights such that $\forall_j w_j \geq 0$ and $\sum w_j = 1$.

The attractive property of GMMs is that they do not require any arbitrary, potentially restrictive, assumptions on the form of an underlying pdf (like e.g. Gaussian assumption). This implies that, compared with parametric approaches, GMMs can adapt to the local geometry of data ensembles (e.g. points distributed in multiple modes or points distributed on a low-dimensional surface in a high-dimensional space) and that they can approximate any continuous density to an arbitrary precision (McLachlan & Peel 2000).

For each estimate of $p(\underline{\mathbf{iv}}|\ C_k)$ is, a GMM was fitted with the method of Figueiredo & Jain (2002) to known indicator values $\underline{\mathbf{iv}}$ of relevés belonging to vegetation type C_k (Matlab code available at www.lx.it.pt/~mtf/mixturecode.zip). The approach of Figueiredo & Jain (2002) is based on the Minimum Message Length (MML) criterion. The rationale behind MML is that if one can built a short code describing one's data that means that one has a good data generation model (Bishop 1995).

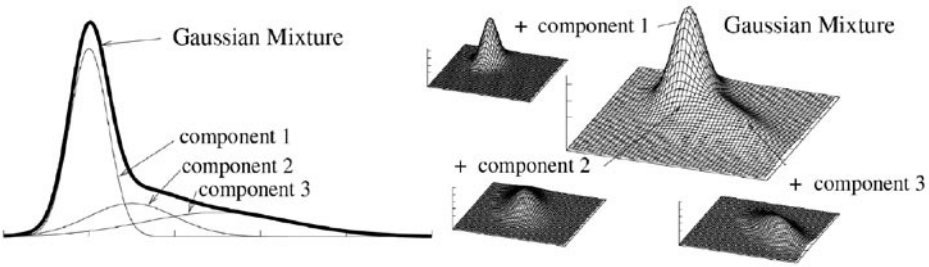

Fig. 2. An example of 1-dimensional (left) and 2-dimensional (right) GMM. In both cases GMM is a linear combination of three pure Gaussian components.

Mathematically, given some data sample $\chi = \{\underline{\mathbf{iv}}_n\}_{n=1}^{N_k}$ the MML criterion for mixture Gaussian pdfs consists of minimizing with respect to $\underline{\boldsymbol{\theta}} \equiv \left[\underline{\Theta}_1,\ldots,\underline{\Theta}_{J_c}, w_1 \ldots w_{J_c}\right]$ the following cost function:

$$\mathcal{L}(\underline{\boldsymbol{\theta}},\chi) = \frac{D_{\underline{\boldsymbol{\theta}}}}{2}\sum_{j=1}^{J_c}\log\left(\frac{N_k w_j}{12}\right) + \frac{J_c}{2}\log\left(\frac{N_k}{12}\right)\frac{DJ_c}{2} - \log(p(\chi\,|\,\underline{\boldsymbol{\theta}})) \quad (4)$$

where $D_{\underline{\boldsymbol{\theta}}} = \dim(\underline{\boldsymbol{\theta}})$ and N_k is the number of sample points which belong to the class C_k. An attractive property of the algorithm of Figueiredo & Jain (2002) is that it is coupled with a selection procedure that automatically determines the number of components J_c. Thus, GMM can be initialized with a relatively large value of J_c, alleviating the need for careful initialization. In this work we initialized means of the Gaussian components in Eq. 3 to 20 randomly chosen data points. The initial covariances $\underline{\underline{\mathbf{C}}}_j$ were made proportional to the identity matrix with the diagonal entries equal to 1/10 of the mean of the variances along each dimension of the data:

$$\sigma^2_{\text{init}} = \frac{1}{10 D_{\text{iv}}}\text{trace}\left(\frac{1}{N_k}\sum_{n=1}^{N_k}\left(\underline{\mathbf{iv}}_n - \underline{\mathbf{m}}\right)\left(\underline{\mathbf{iv}}_n - \underline{\mathbf{m}}\right)^T\right) \quad (5)$$

Where $D_{\underline{\mathbf{iv}}} = \dim(\underline{\mathbf{iv}})$,

$$\underline{\mathbf{m}} = \frac{1}{N_k}\sum_{n=1}^{N_k}\underline{\mathbf{iv}}_n$$

is the global data mean (Figueiredo & Jain 2002). This step was meant to assure the initial density on each data point be reasonably higher than 0.

After obtaining the pdfs of all vegetation types, we can use Eq. 1 to predict the occurrence probability of any vegetation type at known indicator variables. Classification of indicator vector $\underline{\mathbf{iv}}$ into a vegetation type C_k was simply done by taking the vegetation type with the highest posterior occurrence probability:

$$\underset{j\neq k}{\forall} P\left(C_k\,|\,\underline{\mathbf{iv}}\right) > P\left(C_j\,|\,\underline{\mathbf{iv}}\right) \quad (6)$$

It can be shown that this solution is the Bayes decision rule for minimal error (Webb 2002).

Data

For our research we had a database at our disposal with 35 000 relevés taken all over The Netherlands in the period 1928-1988. For the standard work of plant communities in The Netherlands (Schaminée et al. 1995, 1996, 1998; Stortelder et al. 1999), these relevés had already been classified according to their species composition into phytosociological vegetation types using TWINSPAN (Hill 1979) and, to some extent, expert judgement. Only associations – 242 in total – were considered, because at this fundamental hierarchical level of phytosociology, reasonably homogeneous habitat conditions may be expected. A further selection of two grassland groups was made in order to obtain a workable number for this publication:

1. The Dune case: 13 grassland associations, typical of large dune areas in the western part of The Netherlands (3464 relevés).
2. The Pleistocene case: 12 grassland associations that frequently occur in the Pleistocene cover-sand landscape in the eastern part of The Netherlands (2410 relevés).

The selected associations of the Pleistocene case have been used by Grootjans (1985) and by Everts & de Vries (1991) as references describing floristic and ecological characteristics of hydro-ecological gradients. The selection of dune associations is the same as in a predictive model for the Amsterdam Water Supply Dunes (Witte et al. 2006) and based on a vegetation map of this area (van Til & Mourik 1999). Both cases cover associations on a wide variety of soils, ranging from dry to wet, from nutrient-poor to nutrient-rich and from acid to alkaline (see Results for the associations selected). The term 'grassland association' is used here in a very broad sense and includes all non-aquatic and non-woody vegetation, such as: pioneer vegetation, meadows, hay-fields, and heath lands.

A list of ecological species groups that is specifically tailored to The Netherlands was used to determine Indicator Values for all the plant species. This list includes vascular plants (Runhaar et al. 2004; Witte 2002), mosses and liverworts (Dirkse & Kruijsen 1993) and *Characeae* (van Raam & Maier 1993). The derived indicator values

closely resemble the internationally accepted indicator values of Ellenberg (Ellenberg et al. 1992), since both systems distinguish the same important habitat factors: salinity, moisture regime (characterising the availability of both water and oxygen; Runhaar et al. 1997), nutrient availability, and acidity. A major difference, however, is that non-existent combinations of classes have been omitted in the Dutch system (for instance the combination 'saline' and 'nutrient-poor') and that many species have been ascribed to two or more ecological groups, thus taking into account the ecological amplitude of species (Ellenberg's list only mentions the optimum). A complete list of indicator values, including a description of how they were derived, is available as App. 1, 2.

Validation

A randomly chosen half of the secondary database (with C_k and per relevé) was used to calibrate pdfs, while the other half facilitated validation of the classification. To pursue reliable pdfs, only associations with more than 25 relevés were considered. In the validation, *iv*'s were used to predict, per relevé, the Bayesian occurrence probability *P* of vegetation types. Then each relevé was classified to the vegetation type for which the highest *P* was computed (Eq. 6). Finally, to quantify the efficiency of our classification scheme, we constructed a confusion matrix (Kohavi & Provost 1998). Such a matrix contains information about actual and predicted classifications made by a classification system. In other words, each element of the matrix is indexed by the combination of 'observed' vegetation type (rows: classified on the basis of species composition) and 'predicted' vegetation type (columns: classified on the basis of indicator values) the number of relevés. The value of each element determines the number of points belonging to the 'observed' vegetation type that was classified as a 'predicted' vegetation type. Ideally, one hopes for a confusion matrix that consists of only diagonal elements, i.e. elements for which the 'predicted' vegetation type is exactly the same as the 'observed' vegetation type.

Results

Table 1 shows the validation result for ecologically related grassland associations from the Dune landscape in the western part of The Netherlands. On average, 85% of the relevés have been classified correctly based on three indicator values only. For four associations, the three-dimensional space the relevés occupy, with indicator values for moisture regime, nutrient richness and acidity, is presented in Fig. 3A. Fig. 3B gives the density surfaces for these associations. Note that the shape of the density surfaces successfully captures the geometry of the data in Fig. 3A. The validation result

Table 1. Confusion matrix of the Dune case, showing the relationship of the numbers of classified relevés with 'observed' (rows) and 'predicted' (columns) associations. Predicted associations based on indicator values and density functions. # = total number of associations, % = percentage correctly classified relevés. Vegetation codes: 06AC04 = *Samolo-Littorelletum*, 08BB04 = *Typho-Phragmitetum*, 08BD01 = *Cladietum marisci*, 09BA04 = *Junco baltici-Schoenetum nigricantis*, 14AA02 = *Violo-Corynephoretum*, 14BB02 = *Festuco-Galietum veri*, 14CA01 = *Phleo-Tortuletum ruraliformis*, 14CB01 = *Taraxaco-Galietum veri*, 19AA03 = *Botrychio-Polygaletum*, 20AA01 = *Genisto anglicae-Callunetum*, 23AB01 = *Elymo-Ammophiletum*, 27AA02 = *Centaurio-Saginetum*, 28AA01 = *Cicendietum filiformis*. Three associations that were originally selected appeared to have less than 25 relevés for the calibration of reliable pdfs and were therefore excluded from the validation: 06AB01 = *Echinodoro-Potametum graminei*, 09AA01 = *Caricetum trinervi-nigrae* and 09BA03 = *Parnassio-Juncetum atricapilli*.

Observed	Predicted													#	%
	06AC04	08BB04	08BD01	09BA04	14AA02	14BB02	14CA01	14CB01	19AA03	20AA01	23AB01	27AA02	28AA01		
06AC04	**35**	2	4	2	0	0	0	0	0	0	0	1	0	44	80
08BB04	1	**114**	25	0	0	0	0	0	0	0	0	0	4	144	79
08BD01	2	5	**25**	0	0	0	0	0	0	0	0	0	1	33	76
09BA04	4	0	0	**57**	0	0	0	2	1	0	0	8	4	76	75
14AA02	0	0	0	0	**174**	13	13	2	0	0	1	0	0	203	86
14BB02	0	0	0	0	14	**83**	3	9	3	0	3	0	0	115	72
14CA01	0	0	0	0	6	4	**186**	8	0	0	8	0	0	212	88
14CB01	0	0	0	0	4	16	11	**194**	14	0	0	0	0	239	81
19AA03	0	0	0	2	0	0	0	5	**19**	0	0	0	0	26	73
20AA01	0	0	0	0	6	1	0	0	1	**322**	0	0	1	331	97
23AB01	0	0	0	0	1	5	5	1	0	0	**96**	2	0	110	87
27AA02	0	0	0	8	0	0	0	0	0	0	0	**67**	2	77	87
28AA01	5	2	4	9	0	0	0	0	2	0	0	0	**100**	122	82
#	47	123	58	78	205	122	218	221	40	322	108	78	112		
%	74	93	43	73	85	68	85	88	48	100	89	86	89		

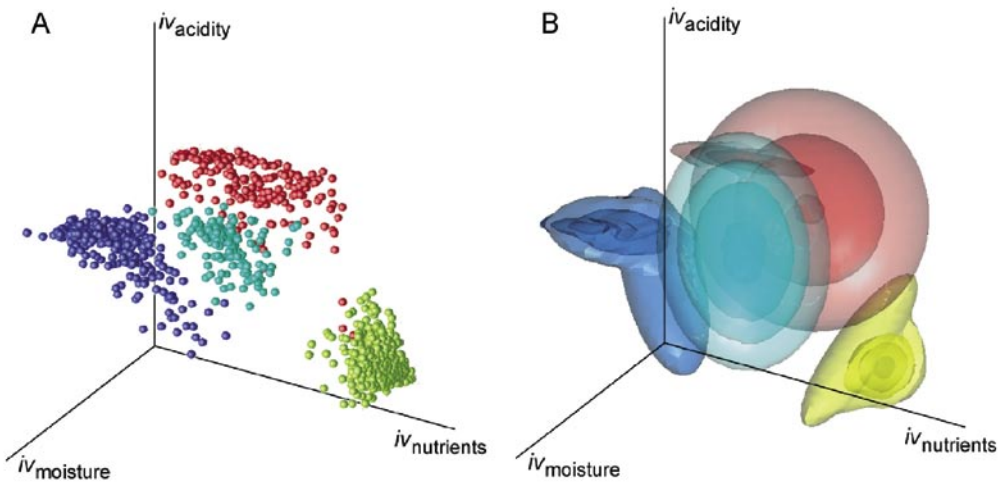

Fig. 3. (A) Three dimensional scatter plot of observed relevés (balls) plotted in relation to their average indicator values iv for 4 associations of the Dune case; colours correspond to different associations (dark blue = 08BB04 *Typho-Phragmitetum*, light blue = 09BA04 *Junco baltici-Schoenetum nigricantis*, green = 20AA01 *Genisto anglicae-Callunetum*, red = 23AB01 *Elymo-Ammophiletum*). (B) For each association, a GMM pdf is fitted through the corresponding cloud of observations. The figure presents isosurfaces of the pdfs, obtained by plotting the values of 0.002, 0.02, 0.1 and 0.5 of the global maximum of each pdf. Note that the association GMM pdfs which induce these isosurfaces capture the geometry of data ensembles from (A). Due to differences in software, the axes of A and B are not exactly the same.

for the Pleistocene case is presented in Table 2. For this case, 83% of the relevés have been classified correctly with the aid of three indicator values.

A closer look at the confusion matrices reveals that misclassified relevés are usually paired with ecologically similar associations. One ready explanation for misclassifications is that more explanatory variables have to be taken into account, such as 'grazing' and 'salinity'. This is, for instance, probably the case for the association 14BB02 *Festuco-Galietum veri* (Table 1), for which occurrence in lime-poor dry dunes depends strongly on grazing with cattle. To incorporate grazing in our method we could use the indicator values for grazing of Briemle & Ellenberg (1994). Besides this, the successional stage may influence classification. For instance, both associations 27AA02 *Centaurio-Saginetum* and 28AA01 *Cicendietum filiformis* are pioneer vegetation types that may be followed in succession by the association 09BA04 *Junco baltici-Schoenetum nigricantis*. As can be seen in Table 1, these associations indeed mix to a certain extent.

Discussion

Correctness of classification

Table 1 and Table 2 demonstrate that is possible to correctly classify vegetation types with the aid of a limited number of indicator values. This good result implicitly proves two important things: (1) the indicator values we employed must have been quite good; (2) the division of the vegetation into associations makes ecological sense: each association occupies a characteristic niche in the three-dimensional space of indicator values. We would never have established a good validation if either of these aspects (indicator values or vegetation division) was poor.

There are many ways to describe the vegetation continuum at the earth's surface by means of an essentially artificial system of vegetation units. Phytosociology is just one of them, which is both popular and criticised (Ewald 2003; Kershaw & Looney 1985; Mueller-Dombois & Ellenberg 1974; Shimwell 1971). Since the classification system forces vegetation samples into discrete units, it is logical that misclassifications occur. This reduces the percentage correctly classified relevés in Tables 1 and 2. An example is the association 16AB01 *Crepido-Juncetum acutiflori* in Table 2, which is syntaxonomically positioned between the associations 16AA01 *Cirsio dissecti-Molinietum* and 16AB04 *Ranunculo-Senecionetum aquatici* (Schaminée et al.

Table 2. Confusion matrix of the Pleistocene case. Meaning of the vegetation codes: 08BC01 = *Caricetum ripariae*, 08BC02 = *Caricetum gracilis*, 09AA03 = *Carici curtae-Agrostietum caninae*, 11AA02 = *Ericetum tetralicis*, 16AA01= *Cirsio dissecti-Molinietum*, 16AB01 = *Crepido-Juncetum acutiflori*, 16AB04 = *Ranunculo-Senecionetum aquatici*, 16BC01 = *Lolio-Cynosuretum*, 19AA01 = *Galio hercynici-Festucetum ovinae*, 19AA02 = *Gentiano pneumonanthes-Nardetum*, 19AA03 = *Botrychio-Polygaletum*, 20AA01 = *Genisto anglicae-Callunetum*. One association that was originally selected appeared to have less than 25 relevés for the calibration of a reliable pdf and was therefore excluded from the validation: 19AA04 = *Betonico-Brachypodietum*.

Observed	Predicted													
	08BC01	08BC02	09AA03	11AA02	16AA01	16AB01	16AB04	16BC01	19AA01	19AA02	19AA03	20AA01	#	%
08BC01	**25**	17	1	0	0	0	3	0	0	0	0	0	46	54
08BC02	12	**24**	0	0	0	0	6	0	0	0	0	0	42	57
09AA03	0	1	**62**	0	5	2	0	0	0	0	0	0	70	89
11AA02	0	0	0	**166**	4	0	0	0	2	10	0	8	190	87
16AA01	0	0	6	0	**114**	10	1	1	1	6	0	0	139	82
16AB01	0	0	1	0	4	**18**	2	1	0	0	0	0	26	69
16AB04	2	0	0	0	0	13	**18**	5	0	0	0	0	38	47
16BC01	0	0	0	0	0	1	6	**185**	0	0	0	0	192	96
19AA01	0	0	0	0	0	0	0	0	**44**	6	0	6	56	79
19AA02	0	0	0	3	4	0	0	0	1	**39**	1	1	49	80
19AA03	0	0	0	0	1	0	0	0	0	0	**25**	0	26	96
20AA01	0	0	0	20	0	0	0	0	36	0	0	**275**	331	83
#	39	42	70	89	32	44	36	192	84	61	26	290		
%	64	57	89	88	86	41	50	96	52	64	96	95		

1995). The results in Table 2 show that the intermediate position of 16AB01 *Crepido-Juncetum acutiflori* leads to misclassification of some relevés.

The occurrence of misclassifications, however, may also be a reason to discuss the division into vegetation types. For example, Table 2 shows that the first two associations (08BC01: *Caricetum ripariae* and 08BC02 *Caricetum gracilis*) are highly intermingled, while one may question whether there is enough ecological or floristic difference to make a distinction between the two (see Schaminée et al. 1995 for a description of their synecology and floristic composition). The decisive floristic difference of these associations, that have no character species, seems to be the dominant *Carex* species (*C. riparia* viz. *C. gracilis*) and this, we guess, is a matter of which of the two species first colonized the soil. For ecological applications, such as predictive models, one might consider merging the two associations into a new vegetation type. This of course improves the confusion matrix: the percentage correctly classified relevés of the merged vegetation type becomes 86% (this was, see last column of Table 2, 54% and 57% for 08BC01 and 08BC02, respectively) and, on average, 85% of the relevés of the Pleistocene case are then classified correctly, instead of 83%.

Of course the results of the validation depend on the selection of vegetation types. To investigate how the selection influences the results, we performed the following analysis: We selected all 242 associations and build probability functions on the basis of four indicator values (Salinity, Moisture regime, Nutrient availability and Acidity). Associations with less than 35 relevés were omitted, thus leaving us with 128 associations for the validation. It appeared that, in this way, 57% of the relevés from the Dune case and 54% of the relevés from the Pleistocene case we classified correctly. Note that we did not make a distinction on the basis of vegetation structure: associations of woodlands, shrubs and grassland were in the same confusion matrix. So, these percentages may be considered as 'worst case' results, in case one would select vegetation types without any ecological knowledge.

Method

The most commonly used method to assess vegetation based on habitat factors is to apply response curves of plant species computed with, for instance, logistic regression or splines from measured habitat factors (e.g. Wamelink et al. 2005), or from indicator values as pseudo habitat factors (e.g. Latour et al. 1994). Compared to this species approach, our method has several advantages. In the first place Bayes' theorem meets the second Kolmogorov axiom, which means that the sum of occurrence probabilities is always 100% (the other approaches do not obey this axiom: the sum may be zero or exceed 100% by far, and how can this be interpreted?). Secondly, the output (vegetation types) can easily be mapped and tested against existing vegetation maps, as was shown in a predictive model by Witte et al. (2006). Finally, the method appears to be rather insensitive to the very selective manner in which relevés have usually been sampled.

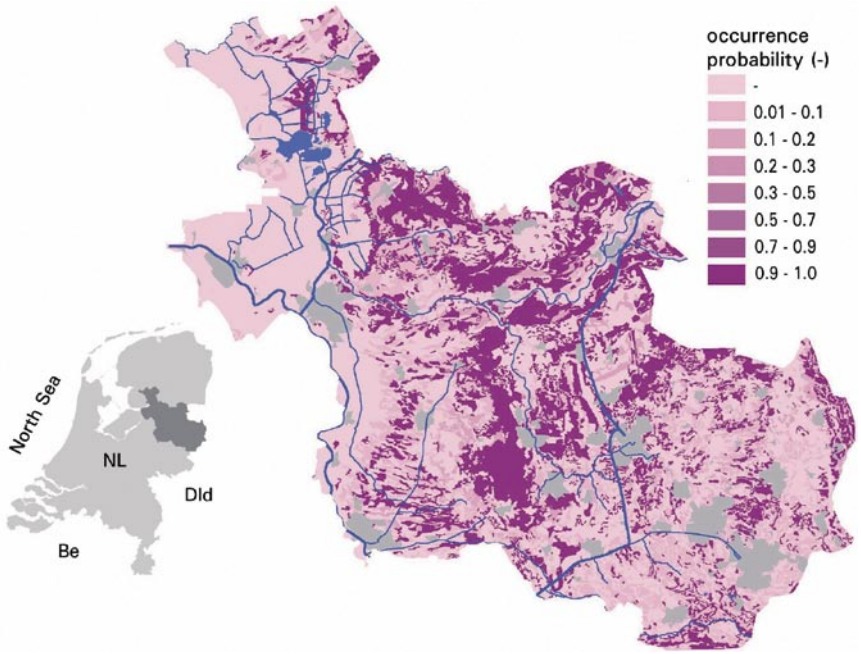

Fig. 4. Potential occurrence probability in the Province of Overijssel (The Netherlands) of the *Genisto anglicae-Callunetum* association (20AA01) (defined by Stortelder et al. 1999; resolution 250 × 250 m; grey = urban area, blue = rivulets), based on current soil types, current groundwater regimes, optimal vegetation management for this vegetation and a historic atmospheric deposition of nitrogen.

Note that our method is very flexible: any vegetation classification system and any list of indicator values can be used to estimate occurrence probability functions. It is even possible to combine different indicator lists (e.g. Ellenberg (1992) for acidity and Landolt (1977) for moisture regime), or to combine indicator lists with published optima (e.g. pH according to Wamelink et al. 2005) or plant traits (e.g. Cornelissen et al. 2003) of plant species.

It should also be noted that 470 000 relevés are available in The Netherlands (www.synbiosys.alterra.nl), which can be used to build density functions of vegetation types. The geographical position of most of these relevés is well known. This enables us to derive regional functions for specific landscapes, such as dunes or brook valleys, thus taking into account regional differences in the species composition of vegetation types and possible regional differences in the indicative value of plant species. Moreover, our method does not depend on a particular division of vegetation into types: any division can be used, provided, of course, that the division makes ecological sense.

Each density function attributes relevés to vegetation types on the basis of plant characteristics that are considered to relate to habitat factors and not, as regular vegetation classification systems do, on the basis of their species composition. This feature makes our method especially suitable to be applied in environmental impact assessment studies, as shown by Witte et al. (2006). As an example, Fig. 4 gives the predicted potential occurrence probability of the *Genisto-Callunetum* association in the province of Overijssel, The Netherlands. Distribution maps of different vegetation types can be combined into one map, showing vegetation types with the highest occurrence probabilities (cf. Eq. 6).

Conclusions

In this paper we introduced a novel method for accurately classifying vegetation types using a limited number of indicator values. Not only does our method sort vegetation types, it also provides an insight into the reliability of the classification by producing occurrence probabilities of vegetation types. The method has much potential for application driven research: it can be used to investigate the quality of indicator values (or, in general, of plant traits) and of vegetation classification systems; it provides information about the ecological requirements of vegetation types, and, finally, it suits predictive vegetation models.

Acknowledgements. This study was carried out in the framework of both the Dutch national research programme Climate Change and Spatial planning (www.klimaatvoorruimte.nl) and the joint research programme of the Dutch Water Utility sector. We thank Peter van Bodegom, Valério Pillar, Otto Wildi, Michael Dale and an anonymous reviewer for their valuable comments on the manuscript.

References

Bishop, M.C. 1995. *Neural networks for pattern recognition*. Oxford University Press, New York, NY, US.

Briemle, G. & Ellenberg, H. 1994. Zur Mahdverträglichkeit von Grünlandpflanzen. *Natur Landschaft* 69: 139-147.

Cornelissen, J.H.C., Lavorel, S., Garnier, E., Díaz, S., Buchmann, N., Gurevich, D.E., Reich, P.B., ter Steege, H., Morgan, H.D., van der Heijden, M.G.A., Pausas, J.G. & Poorter, H. 2003. A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. *Aust. J. Bot.* 51: 335-380.

Diekmann, M. 2003. Species indicator values as an important tool in applied ecology – a review. *Basic Appl. Ecol.* 4: 493-506.

Dirkse, G.M. & Kruijsen, B.W.J.M. 1993. Indeling in ecologische groepen van Nederlandse blad- en levermossen. *Gorteria* 19: 1-29.

Dzwonko, Z. 2001. Assessment of light and soil conditions in ancient and recent woodlands by Ellenberg indicator values. *J. Appl. Ecol.* 38: 942-951.

Ellenberg, H., Weber, H.E., Düll, R., Wirth, V., Werner, W. & Paulißen, D. (eds.) 1992. *Zeigerwerte von Pflanzen in Mitteleuropa*, 3. Aufl. Scripta Geobotanica 18. Verlag Erich Goltze, Göttingen, DE.

Everts, F.H. & de Vries, N.P.J. 1991. *De vegetatieontwikkeling van beekdalsystemen: een landschapsoecologische studie van enkele Drentse beekdalen*. Ph.D. Thesis, Groningen, NL.

Ewald, J. 2003. A critique for phytosociology. *J. Veg. Sci.* 14: 291-296.

Figueiredo, A.T. & Jain, A.K. 2002. Unsupervised Learning of Finite Mixture Models. *IEEE Transactions of pattern analysis and machine intelligence* 24: 381-396.

Grootjans, A.P. 1985. *Changes of groundwater regime in wet meadows*. Ph.D. Thesis, Groningen, NL.

Hill, M.O. 1979. *TWINSPAN - A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes*. Cornell University, Ithaca, NY, US.

Käfer, J. & Witte, J.P.M. 2004. Cover-weighted averaging of indicator values in vegetation analyses. *J. Veg. Sci.* 15: 647-652.

Kershaw, K.A. & Looney, J.H.H. 1985. *Quantitative and dynamic plant ecology*, 3rd ed. Arnold, London, UK.

Kohavi, R. & Provost, F. 1998. *Machine Learning* 30: 271-274.

Landolt, E. 1977. *Ökologische Zeigerwerte zur Schweizer Flora*. Veröffentlichungen des Geobotanischen Institutes der Eidg. Techn. Hochschule, no. 64. Stiftung Rübel, Zürich, CH.

Latour, J.B., Reiling, R. & Slooff, W. 1994. Ecological standards for eutrophication and desiccation: perspectives for a risk assessment. *Water Air Soil Pollut.* 78: 265-279.

McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. 2006. Rebuilding community ecology from functional traits. *Trends Ecol. Evol.* 21: 178-185.

McLachlan, G. & Peel, D.A. 2000. *Finite mixture models*. Wiley Interscience, New York, NY, US.

Mueller-Dombois, D. & Ellenberg, H. 1974. *Aims and methods of vegetation ecology*. Wiley, New York, NY, US.

Runhaar, J., Van Landuyt, W., Groen, C.L.G., Weeda, E.J. & Verloove, F. 2004. Herziening van de indeling in ecologische soortengroepen voor Nederland en Vlaanderen. *Gorteria* 30: 12-26.

Runhaar, J., Witte, J.P.M. & Verburg, P.H. 1997. Groundwater level, moisture supply and vegetation. *Wetlands* 17: 528-538.

Schaffers, A.P. & Sýkora, K.V. 2000. Reliability of Ellenberg indicator values for moisture, nitrogen and soil reaction: a comparison with field measurements. *J. Veg. Sci.* 11: 225-244.

Schaminée, J.H.J., Stortelder, A.H.F. & Westhoff, V. 1995. *De vegetatie van Nederland*. Vol. 2. Opulus Press, Uppsala, SE.

Schaminée, J.H.J., Stortelder, A.H.F. & Weeda, E.J. 1996. *De vegetatie van Nederland*. Vol. 3. Opulus Press, Uppsala, SE.

Schaminée, J.H.J., Weeda, E.J. & Westhoff, V. 1998. *De vegetatie van Nederland*. Vol. 4. Opulus Press, Uppsala, SE.

Schmidtlein, S. 2005. Imaging spectroscopy as a tool for mapping Ellenberg indicator values. *J. Appl. Ecol.* 42: 966-974.

Shimwell, D.W. 1971. *The description and classification of vegetation*. Sidgwick & Jackson Ltd., London, UK.

Stortelder, A.H.F., Schaminée, J.H.J. & Hommel, P.W.F.M. 1999. *De vegetatie van Nederland*. Vol. 5. Opulus Press, Uppsala, SE.

van Raam, J.C. & Maier, E.X. 1993. Overzicht van de Nederlandse Kranswieren. *Gorteria* 18: 111-116.

van Til, M. & Mourik, J. 1999. *Hiërogliefen van het zand: vegetatie en landschap van de Amsterdamse waterleidingduinen*. Gemeentewaterleidingen Amsterdam, Amsterdam, NL.

Wamelink, G.W.W., Goedhart, P.W., van Dobben, H.F. & Berendse, F. 2005. Plant species as predictors of soil pH: Replacing expert judgement with measurements. *J. Veg. Sci.* 16: 461-470.

Webb, A.R. 2002. *Statistical pattern recognition*. 2nd ed. John Wiley & Sons Ltd, London, UK.

Witte, J.P.M. 2002. The descriptive capacity of ecological plant species groups. *Plant Ecol.* 162: 199-213.

Witte, J.P.M., de Haan, M., Raterman, B. & Aggenbach, C. 2006. *PROBE — Versie 1: effecten van grondwaterbeheer, atmosferische depositie, maaien en plaggen*. Kiwa Water Research, Nieuwegein, NL.

Witte, J.P.M., Meuleman, J.A.M., van der Schaaf, S. & Raterman, B. 2004. Eco-hydrology and bio-diversity. In: Feddes, R.A., de Rooij, G.H. & van Dam, J.C. (eds.) *Unsaturated zone modelling: Progress, challenges and applications*, pp. 301-329. Kluwer, Dordrecht, NL.

Received 1 August 2006;
Accepted 26 January 2007;
Co-ordinating Editor: V.D. Pillar.

For App. 1-2, see also JVS/AVS Electronic Archives; www.opuluspress.se/